

MATH1905 Statistics

Lecture 11

Lecturer: Marc Raimondo
Carlaw 817

Lectures: Mon.11am (Chem LT1), Tue.8am (Chem LT1)

Student consultations: Tuesday 2-3pm

General information, tutes, solutions etc...

<http://www.maths.usyd.edu.au/u/UG/JM/MATH1905/>

or

First Year Office (FYO), Carlaw 520.

Continuous random variables

When the outcome of a random experiment is a measurement, this outcome itself can be taken to be the associated rv, X . In this case X is a variable which can take uncountably many values, x denoting a typical value. p e.g. measuring height, weight on a continuous scale... The only sensible value which can be given to $P(X = x)$ is zero, so that we need a different approach to that used for discrete rv's.

Whereas the probability distribution $\{p_i\}$ was the basic object in the discrete case, in the continuous case that role is played by the 'probability density function'.

Distribution function

This is defined for the continuous rv X , as in the discrete case, by

$$F(x) = \mathbf{P}(X \leq x)$$

but the argument x is now continuous.

In the discrete case a distribution function is a STEP function

In the continuous case a distribution function is a SMOOTH function

Properties of the distribution function of a continuous rv X

$$F(x) = \mathbf{P}(X \leq x) = \mathbf{P}(X < x),$$

$$\mathbf{P}(X > x) = 1 - F(x),$$

$$0 \leq F(x) \leq 1$$

$$\text{If } a < b \text{ then } \mathbf{P}(a < X < b) = F(b) - F(a).$$

$$\text{If } a < b \text{ then } F(a) \leq F(b)$$

Proof

Recall that if $A \subset B$ then $P(A) < P(B)$

$$a < b, \{X < a\} \subset \{X < b\}$$

$$P(\{X < a\}) < P(\{X < b\}) \quad \bullet$$

$$\{X < b\} = \{X \leq a\} \cup \{X > a \cap X < b\} \quad m.e$$

$$P(X < b) = P(X \leq a) + P(a < X < b)$$

$$P(a < X < b) = F(b) - F(a) \quad \bullet$$

$$P(X > x) = 1 - P(\{X > x\}^c) = 1 - P(X \leq x) \quad \bullet$$

Example The rv X has distribution function $F(x) = 0$ for $x < 0$; $F(x) = x^{1/2}$ for $0 \leq x \leq 1$ and $F(x) = 1$ for $x > 1$. Find $\mathbf{P}(\frac{1}{2} < X < \frac{2}{3})$.

$$\mathbf{P}(\frac{1}{2} < X < \frac{2}{3}) = F(\frac{2}{3}) - F(\frac{1}{2}) = \sqrt{\frac{2}{3}} - \sqrt{\frac{1}{2}} \approx \mathbf{0.11}$$

Probability density function

The expression $\frac{F(x+h) - F(x)}{h}$ can be interpreted as the average 'density of probability' over the interval $(x, x + h)$. Letting $h \rightarrow 0$ gives the **probability density function**:

$$f(x) = \frac{d}{dx}F(x).$$

Since a function is the integral of its derivative,

$$\mathbf{P}(a < X < b) = \int_a^b f(x) dx.$$

In our example: $f(x) = \frac{\partial}{\partial x}(\sqrt{x}) = \frac{1}{2}x^{-\frac{1}{2}}, 0 < x < 1.$

Properties of probability density function

For any pdf f ,

- $f(x) \geq 0$ for all x ,
- $\int_{-\infty}^{\infty} f(x) dx = 1$

- If X has pdf f , then for any region A ,
 $\mathbf{P}(X \in A) = \int_A f(x) dx$

- If X has pdf f , then for small δ , $\mathbf{P}(x < X < x + \delta) \approx \delta f(x)$

- $F(x) = \int_{-\infty}^x f(y) dy$

Example The rv X has pdf f given by

$$f(x) = cxe^{-\lambda x^2} \text{ for } x \geq 0; 0 \text{ for } x < 0$$

(λ is a positive constant).

- (i) What value does c take?
- (ii) Find the distribution function of X .
- (iii) Calculate $\mathbf{P}\left(\frac{1}{2\sqrt{\lambda}} < X < \frac{1}{\sqrt{\lambda}}\right)$.
- (iv) Sketch the pdf.

Solution i) We find c by solving $\int f(x) dx = 1$.

$$\int_0^{\infty} c x e^{-\lambda x^2} dx = 1$$

we change variable in $\int_0^{\infty} x e^{-\lambda x^2} dx$

$$u = x^2, du = 2x dx, x = \sqrt{u}, du = 2\sqrt{u} dx$$

$$\begin{aligned} \int_0^{\infty} x e^{-\lambda x^2} dx &= \int_0^{\infty} \sqrt{u} e^{-\lambda u} du / (2\sqrt{u}) = \frac{1}{2} \left[\frac{e^{-\lambda u}}{-\lambda} \right]_0^{\infty} = \\ &= - \left(-\frac{1}{2\lambda} \right) = \frac{1}{2\lambda} \end{aligned}$$

Thus $c \times \frac{1}{2\lambda} = 1$ gives $c = 2\lambda$

The density is $f(x) = 2\lambda x e^{-\lambda x^2}, x \geq 0$.

Solution ii) The distribution function is

$$F(x) = \int_0^x 2\lambda u e^{-\lambda u^2} du = 2\lambda \int_0^x u e^{-\lambda u^2} du$$

A similar change of variable as before ($v = u^2$) gives

$$F(x) = 2\lambda \frac{1}{2} \left[\frac{e^{-\lambda v}}{-\lambda} \right]_0^{x^2} = 1 - e^{-\lambda x^2}, \quad x \geq 0.$$

$$\text{iii) } P\left(\frac{1}{2\sqrt{\lambda}} < X < \frac{1}{\sqrt{\lambda}}\right) = F\left(\frac{1}{2\sqrt{\lambda}}\right) - F\left(\frac{1}{\sqrt{\lambda}}\right)$$

$$= (1 - e^{-\lambda \frac{1}{\lambda}}) - (1 - e^{-\lambda \frac{1}{4\lambda}}) = e^{-\frac{1}{4}} - e^{-1} \approx 0.41$$

Mean and variance

These are defined by analogy with discrete rv's:

$$\mu = \mathbf{E}(X) = \int x f(x) dx,$$

$$\sigma^2 = \mathbf{E}(X - \mu)^2 = \int (x - \mu)^2 f(x) dx$$

$$= \int x^2 f(x) dx - \mu^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2$$

Example: $F(x) = \sqrt{x}$, $0 \leq x \leq 1$.

The density is $f(x) = \frac{\partial}{\partial x}(\sqrt{x}) = \frac{1}{2}x^{-\frac{1}{2}}$, $0 < x < 1$.

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \frac{1}{2}x^{-\frac{1}{2}} dx = \int_0^1 \frac{1}{2} x^{\frac{1}{2}} dx$$

$$E(X) = \frac{1}{2} \left[\frac{x^{\frac{3}{2}}}{\frac{3}{2}} \right]_0^1 = \frac{1}{2} \left(\frac{2}{3} - 0 \right) = \frac{1}{3}$$

$$E(X^2) = \int_0^1 x^2 \frac{1}{2}x^{-\frac{1}{2}} dx = \int_0^1 \frac{1}{2} x^{\frac{3}{2}} dx$$

$$E(X^2) = \frac{1}{2} \left[\frac{x^{\frac{5}{2}}}{\frac{5}{2}} \right]_0^1 = \frac{1}{2} \left(\frac{2}{5} - 0 \right) = \frac{1}{5}$$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{1}{5} - \left(\frac{1}{3}\right)^2 = \frac{4}{45}$$

Chebyshev's inequality

If a rv X has mean μ and variance σ^2 , then for any positive number c ,

$$P(|X - \mu| \geq c\sigma) \leq 1/c^2$$

Proof (continuous case)

$$I = \{|X - \mu| \geq c\sigma\} = \{x : |x - \mu| \geq c\sigma\} = \{x : \frac{|x - \mu|}{c\sigma} \geq 1\}$$

$$P(|X - \mu| \geq c\sigma) = \int_I f(x)dx \leq \int_I \frac{|x - \mu|}{c\sigma} f(x)dx$$

$$P(|X - \mu| \geq c\sigma) \leq \int \frac{(x - \mu)^2}{c^2\sigma^2} f(x)dx = \frac{1}{c^2}$$

R commands for density functions

Normal $Z \sim N(0, 1)$, $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$: `dnorm(x)`

R commands for distribution functions

Normal $Z \sim N(0, 1)$

$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \phi(x)dx$: `pnorm(x)`

Lecture 12: Normal random variables

This is the most important example of a continuous rv. It is used to model many physical and biological measurements, and also arises as an approximation to a sum of independent rv's.

THE STANDARD NORMAL VARIABLE A standard result from calculus shows that the function

$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad -\infty < x < \infty$$

is a pdf.

The rv Z with pdf ϕ is said to be a *standard normal* variable.

Standard normal and useful identities

This rv is so important that special letters are usually reserved for the rv (Z), for its pdf (ϕ) and for its distribution function (Φ).

$$\mathbf{P}(Z \leq z) = \Phi(z) = \int_{-\infty}^z \phi(x) dx, \quad -\infty < z < \infty$$

- $\phi(-x) = \phi(x)$ (Symmetry w.r.t. 0)
- $\Phi(-x) = 1 - \Phi(x)$ and $\mathbf{P}(|Z| \leq x) = 2\Phi(x) - 1$.

Proof. $\mathbf{P}(|Z| \leq x) = 1 - \mathbf{P}(|Z| \geq x) = 1 - \mathbf{P}(Z > x \cup Z < -x)_{m.e.} = 1 - \mathbf{P}(Z > x) - \mathbf{P}(Z < -x) =_{sym} 1 - 2\mathbf{P}(Z > x) = 1 - 2(1 - \mathbf{P}(Z \leq x)) = 1 - 2 + 2\Phi(x) = 2\Phi(x) - 1$

The Standard normal table—and R. No explicit closed-form expression is available for $\Phi(x)$, but $\Phi(x)$ is available in R.

Example: Calculate

1. $P(Z < 1.5) = pnorm(1.5) = 0.9332$

2. $P(Z > -0.8) = 1 - P(Z \leq -0.8) = 1 - pnorm(-0.8) = 0.7881$

3. $P(-1 < Z \leq 2) = P(Z \leq 2) - P(Z \leq -1) = pnorm(2) - pnorm(-1) = 0.9772 - 0.2420 = 0.7352$

4. $P(|Z| \leq 1.96) = 2P(Z \leq 1.96) - 1 = 2pnorm(1.96) - 1 = 2 \times 0.9750 - 1 = 0.9500$

Standardized rv's

If X is an rv with mean μ and variance $\sigma^2 > 0$ then

$$Y = \frac{X - \mu}{\sigma}$$

is called the standardized version of X .

Fact $E(Y) = 0$ and $var(Y) = 1$.

Proof. $E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}(E(X) - \mu) = \frac{1}{\sigma}0 = 0$

$$E\left(\frac{X - \mu}{\sigma}\right)^2 = \frac{E(X - \mu)^2}{\sigma^2} = 1$$

Fact The standard normal rv Z has $E(Z) = 0$ and $var(Z) = 1$.

We write $Z \sim \mathcal{N}(0, 1)$.

Example Suppose X is uniform on (a, b) . ($a < b$). Find the values of a and b for which X is standardized. **Solution.** The uniform density is constant ($= c$) over the interval (a, b) . The value of c is chosen so that the area of the rectangle is 1. Here $c = \frac{1}{b-a}$.

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}$$

So $E(X) = 0$ if and only if $a = -b$

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b = \frac{b^3 - a^3}{3(b-a)}$$

$$\text{If } a = -b, E(X^2) = \frac{b^3 - (-b^3)}{(2b) \times 3} = \frac{b^2}{3}$$

So $Var(x) = E(X^2) = 1$ if and only if $b = \sqrt{3}$

The general normal variable

The normal variable X with mean μ and variance $\sigma^2 > 0$ (we write $X \sim \mathcal{N}(\mu, \sigma^2)$) is defined as the variable which, when standardized, is $\mathcal{N}(0, 1)$, i.e. $\frac{X-\mu}{\sigma} = Z$.

Fact If $X \sim \mathcal{N}(\mu, \sigma^2)$ then its distribution function is

$$F(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Proof. $Y = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$, $P(X \leq x) = P(\sigma Y + \mu \leq x) = P(Y \leq \frac{x-\mu}{\sigma}) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

This simple relation between $F(x)$ and $\Phi(x)$ is the reason why only the 'standard normal table' is (/was) needed...

Example...In R • If $X \sim \mathcal{N}(5, 4^2)$, find

1. $P(X < 12) = pnorm(\frac{12-5}{4}) = 0.96$ (2dp)

Alternatively : $P(X < 12) = pnorm(12, 5, 4) = 0.96$ (2dp)

2. $P(X > 0) = 1 - P(X \leq 0) = 1 - pnorm(0, 5, 4) \approx 0.89$

3. $P(-1 < X < 6) = pnorm(6, 5, 4) - pnorm(-1, 5, 4) = 0.53$

4. $P(-2 \leq X \leq 12) = \dots$

Example A machine is set to produce items of diameter 85 units, but because of random variations the distribution of diameters produced is described by a normal variable

$$X \sim \mathcal{N}(85, 0.8^2).$$

In a day's production of 2000 items, about how many items will have diameters

1. exceeding 85.6 units? $Y = \text{number of items with Diameter} > 85.6$

$Y \sim B(2000, p)$ where $p =$ the probability that one item will have diameter exceeding 85.6 units. $p = P(X > 85.6) = 1 - P(X \leq 85.6) = 1 - \text{pnorm}(85.6, 85, 0.8) = 0.226$ In one day there will be (on average) about $E(Y) = 2000 \times 0.226 \approx 453$ such items.

2. not exceeding 83 units? The probability that one item will have diameter not exceeding 83 units is $P(X < 83) = \text{pnorm}(83, 85, 0.8) = 0.0062$ In one day there will be (on average) about $E(Y) = 2000 \times 0.0062 \approx 12$ such items.
3. lying between 84.6 and 85.4 units? ...765 such items

The quantile function in R

The quantile function is the *inverse* of the distribution function...that is if $\alpha, 0 < \alpha < 1$ is a fixed number. The quantile of order α of a rv X is the number x_α such that

$$P(X \leq x_\alpha) = \alpha.$$

In R the `qnorm()` function gives quantiles for the standard normal, for example:

$$\text{qnorm}(0.95) = 1.645 \text{ (3dp)}$$

you can check

$$\text{pnorm}(1.645) = 0.95 \text{ (3dp)}$$